Update on the HDF5 standardization effort

Elena Pourmal, Mike Folk

The HDF Group

July 20, 2006

SPG meeting, Palisades, NY

Outline

- HDF5 status
- Lessons learned or thoughts about the standardization process

HDF5 Status

- Three documents were submitted to SPG in March 2006
  - HDF5 Data Model
  - HDF5 File Format (release 1.6.5)
  - HDF5 Reference Manual (release 1.6.5)
- Current response from reviewers (4 total, one is for HDF4)
  - Reviews emphasized
    - HDF complexity
    - Backward-forward compatibility
  - No reviews on “accuracy” and “clarity”, mostly address “usefulness” of HDF

Struggles with HDF5 standardization

- HDF5 is represented at least by 4 layers
    1. Abstract Data Model
    2. APIs
    3. I/O library
    4. File Format (XML, binary)
- Should these be standardized independently, or are they all of a piece?

Struggles with HDF5 standardization

- In our first attempt, we treated them of a piece
  - Linked #1 & #4: storage layout treated as part of data model
  - To an extent #2 also linked: object methods reflected by APIs
- But one layer can evolve without changes in another
  - E.g. variable size chunking will need file format change, but it will not change the data model
  - E.g. new compression will tweak APIs, but may not change format or data model
- Compare to, say, OPeNDAP
  - Just one layer involved -- doesn't describe persistent storage

Struggles with HDF5 standardization

- Objects “in memory” vs. objects “in a file”
  - May lead to different implementation
- Terminology usage (e.g. “persistent” object)
  - Document is not always clear and accurate
- In our first attempt:
  - We didn’t describe objects in memory
  - Removed “persistent” to make document “clear” and introduces inaccuracy: only objects stored in a file (persistent objects) may have attributes

Thoughts about the standardization process

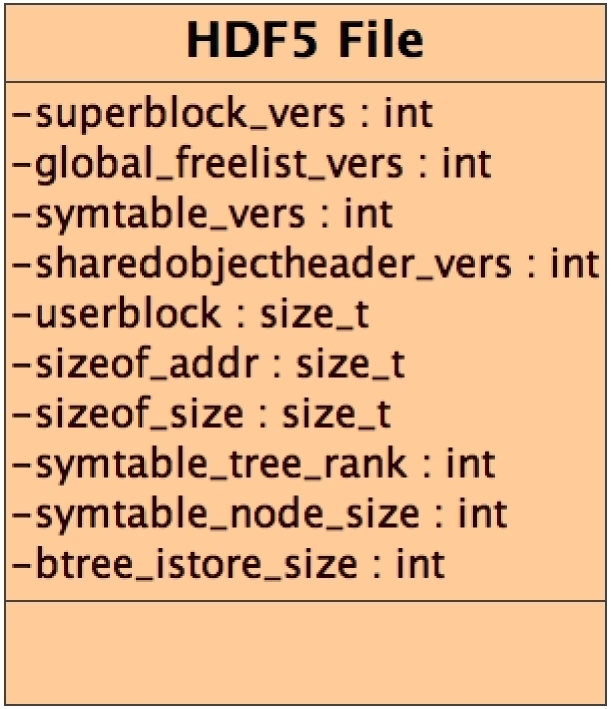
- **It's hard!!**
  - Takes a lot of work to write or review the documents
- **Can we spread out the work?**
  - Assign different parts of the doc to different people
  - Different people may address different issues
  - Different criteria for different reviewers
    - accuracy vs. usefulness
  - But someone still needs to review the whole thing
  - And include a technical writer with special knowledge

Thoughts about the standardization process

- Iterative approach definitely the way to go
- Both standard and review templates were very useful in our work
- Recommend common documentation formats
  - Usage of UML, for example

Example: File Class Diagram “memory” representation

- Concise view
- Easy to find errors
- Easy to review

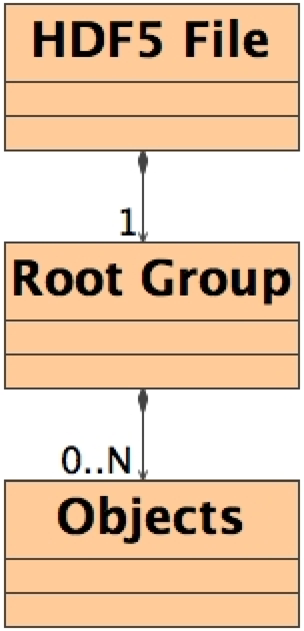


HDF5 File

-superblock_vers : int
-global_freelist_vers : int
-symtable_vers : int
-sharedobjectheader_vers : int
-userblock : size_t
-sizeof_addr : size_t
-sizeof_size : size_t
-symtable_tree_rank : int
-symtable_node_size : int
-btree_istore_size : int

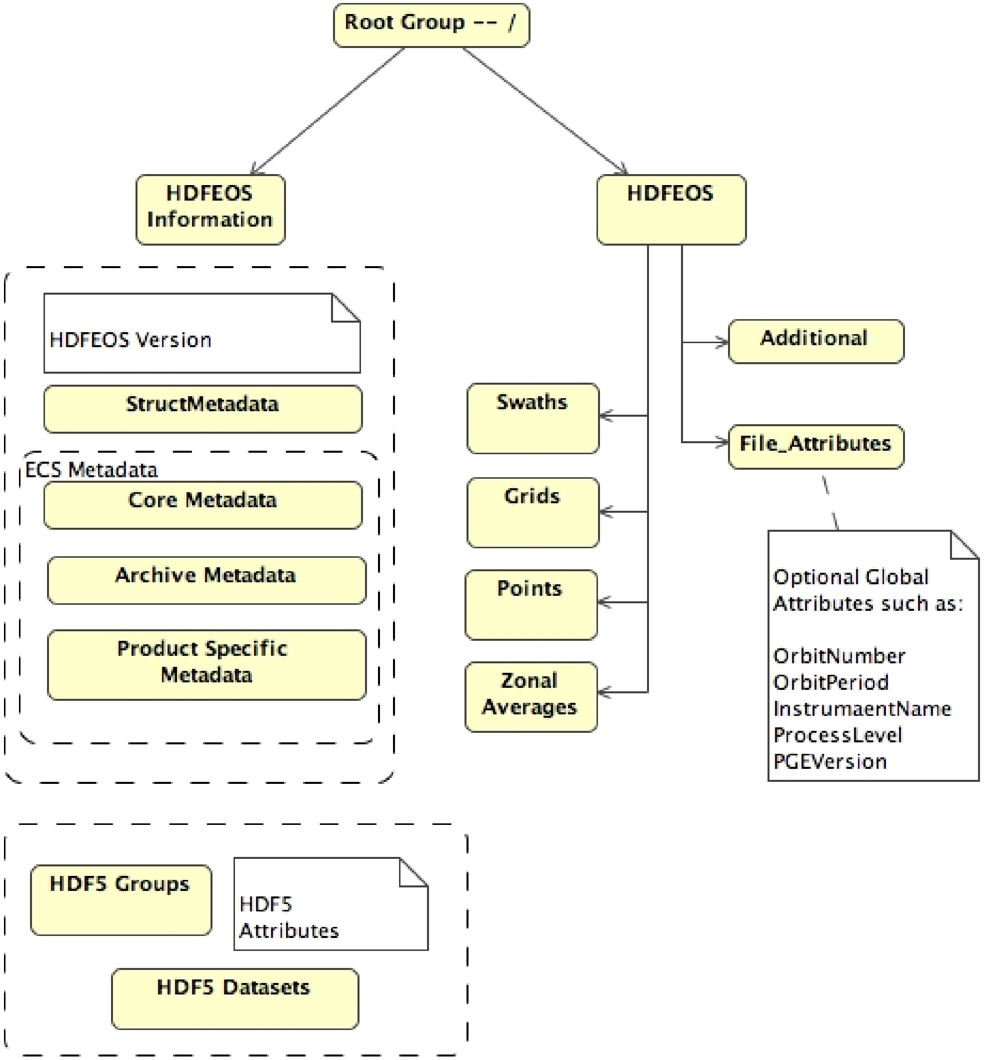
Example: HDF5 File, Root Group, and Objects class diagram

- Shows associations
- Easy to understand the model



HDF-EOS before …

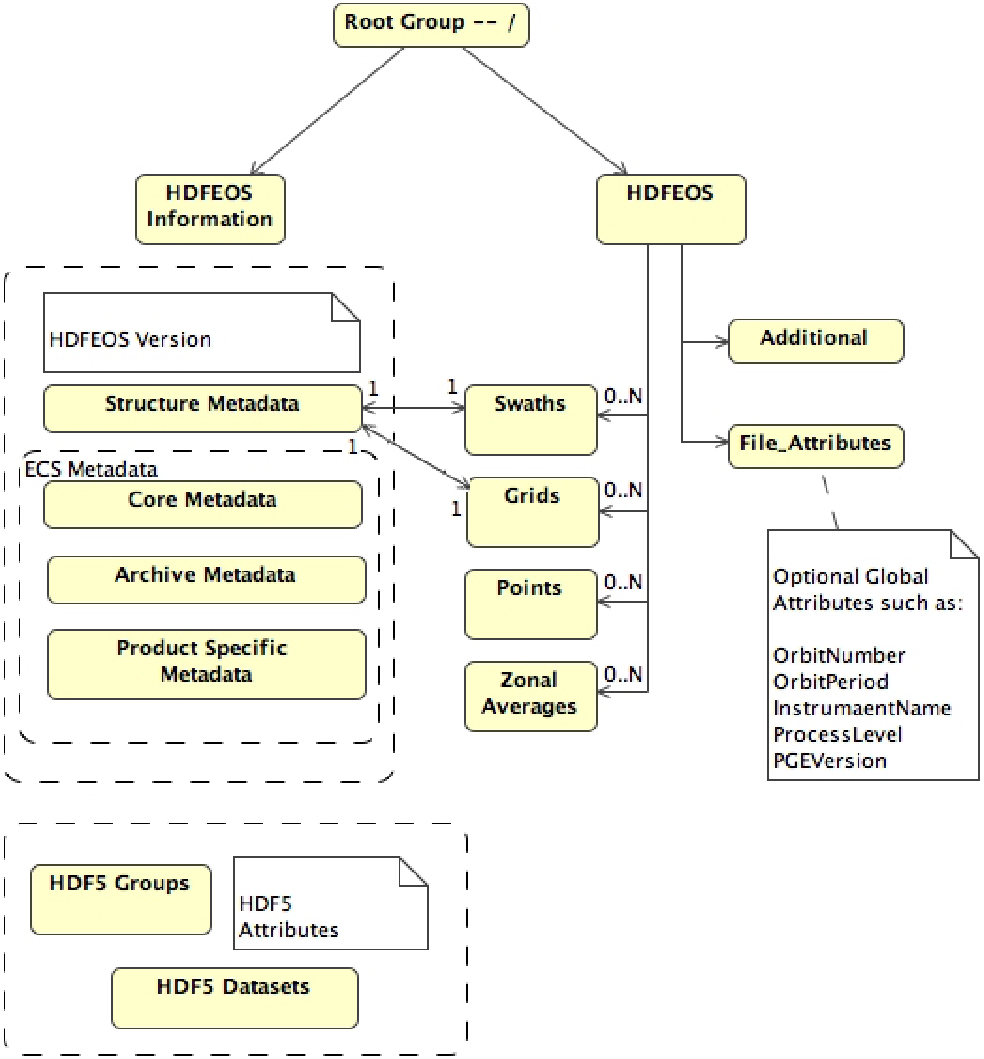
association between objects is missing



HDF-EOS and after

with association shown

Slide has an error. Can you find it ☺?



IETF a good model in many ways, but…

- Consider who participates in IETF
  - Mainly lots of developers
  - Technologies tend to be near and dear to their hearts
  - People excited to participate, volunteer
  - Don’t mind spending lots of time on the topics
  - Often funded by employer to participate
  - And how many IETF standards die on the vine?

IETF a good model, but…

- Vs. who participates in ES-DSWG
    - Earth Scientists?
        - The purpose of the HDF-EOS was to shield them from worrying what is going on under the hood
        - Now we ask them review details they would prefer not to know
        - So have them assess usefulness, but not accuracy
    - IT folks and others
        - Definitely appropriate, but don’t expect the passion IETF generates

IETF vs. Earth Science standards

- IETF standards often much less complex than ES standards
- Some ISO standards perhaps a better model for ES standards
    - E.g. EXPRESS/STEP more like HDF-EOS 5 than like TCP. Complex, multi-faceted, domain-related
    - Standardization more resource-intensive than IETF
    - Participation often supported by employer, can be full-time

Different standards for different goals

- Why do standardization? What are our goals?
  - Sharing: To share data and tools
  - Access: To make data more readily available
  - Integrity: To use data in an appropriate or predefined way
  - Preservation: To be able to understand and use data in future
  - Others?
- Each goal achieved by different layers of our standards
  - OPeNDAP – sharing and access
  - HDF-EOS5 model and API – sharing, integrity, preservation
  - HDF5 data model – sharing
  - HDF5 File Format -- preservation
- Some goals may be achieved just by one layer
  - E.g. MATLAB needs just HDF-EOS5 API to *access* EOS data

One other observation

- How about leveraging HDF5 standardization effort with other usage of HDF5 within NASA
  - CGNS
  - NetCDF4
  - Others?